

Incorporating prior knowledge in environmental sampling: ranked set sampling and other double sampling procedures

Nicolle A. Mode¹, Loveday L. Conquest^{2,*†} and David A. Marker³

¹*Quantitative Ecology and Resource Management, University of Washington, Seattle, WA 98195, U.S.A.*

²*Fisheries/Quantitative Science, University of Washington, Box 355020, Seattle, WA 98195, U.S.A.*

³*Westat, Inc., 1650 Research Blvd., Rockville, MD 20850, U.S.A.*

SUMMARY

Environmental sampling can be difficult and expensive to carry out. Those taking the samples would like to integrate their knowledge of the system or their judgment about the system into the sample selection process to decrease the number of necessary samples. However, mere convenience or non-random sampling can severely limit statistical inference. Methods do exist that integrate prior knowledge into a random sampling procedure that allows for valid statistical inference. Double sampling methods use this extra information to select samples for measurement, thus reducing the number of necessary samples (in order to achieve a desired objective) and thereby reducing sampling costs. The level of prior information required can range from a linear relationship with a known auxiliary variable to simple ranking based on auxiliary information. We examine three types of double sampling methods (ranked set sampling, weighted double sampling and double sampling with ratio estimation), with accompanying examples from Oregon stream habitat data. All three methods can provide increased precision and/or lower sampling costs over simple random sampling. The appropriate double sampling method for the data and research situation depends upon the type of prior information available. The categories of prior information are summarized in a table and illustrated using the example data. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: ranked set sampling; double sampling; two-phase sampling; simple random sampling; relative precision; prior knowledge

1. INTRODUCTION

Many sampling methods are available to researchers that lower sampling costs by using auxiliary information outside the main investigation. However, such a cost-reducing or double sampling method is useful to researchers only if they know when and how to properly use it. This article examines how the degree of prior knowledge about the variables being measured can help determine when a sampling method is appropriate. Specifically, several sampling methods will be compared which use frugal information obtained from a first phase of sampling to obtain a more precise estimate from a costly

*Correspondence to: Loveday L. Conquest, Fisheries/Quantitative Science, University of Washington, Box 355020, Seattle, WA 98195, U.S.A.

†E-mail: conquest@u.washington.edu

Contract/grant sponsor: U.S. Environmental Protection Agency; contract/grant numbers: CR825173-01-0 & CR824682.

second phase of sampling. Prior knowledge can include information about the distribution of the variable of interest, or about the relationship between the variable of interest and auxiliary information. The article is organized as follows. Section 2 reviews three double sampling methods, while Section 3 explains the concept of prior knowledge associated with these methods. In Section 4, the appropriateness of these sampling methods relative to prior knowledge about the variables is illustrated on a sample data set.

2. DOUBLE SAMPLING

Environmental sampling can be difficult and expensive to implement. Many researchers want to integrate prior knowledge (e.g. about the spatial distribution of a soil contaminant, or about the distribution of habitat areas in a stream) into the selection process to decrease the number of necessary samples. However, convenience or non-random sampling can bias results and limit statistical inference. Alternative methods do exist that integrate prior or extra knowledge into a random sampling method which allow for statistical inference. Double sampling methods use this prior or extra information from frugal sampling to select samples for more costly measurement. These methods reduce the number of necessary samples and thereby reduce sampling costs. The level of extra or prior information required can range from a linear relationship with a known auxiliary variable to simple ranking based on auxiliary information.

Three types of double sampling methods are examined in this article, with each introduced below: ranked set sampling, weighted double sampling, and double sampling with ratio estimation.

2.1. Ranked set sampling

In ranked set sampling (RSS: McIntyre, 1952; Takahasi and Wakimoto, 1968) the extra information is a frugal measurement which adds information in the form of ranked sets of data. Small sets of samples are ranked using a frugal measurement, and subsequently one sample from each set is measured using the actual and more costly measurement. Generally, RSS involves an initial ranking of n samples of size n by way of the frugal measurement. Following this, the researcher uses a costly measurement to observe the first order statistic (smallest observation) from the first sample, the second order statistic (second smallest observation) from the second sample, and so on, until the n th order statistic from the n th sample yields a final sample of size n from the initial n^2 observations. Large sample sizes can be obtained by repeating the process on m cycles. This approach keeps the set size small to limit ranking errors. Repeating the process m times yields a final sample of size nm from an initial n^2m observations.

2.2. Weighted double sampling

Weighted double sampling integrates frugal information into the sample selection process by categorizing the samples into groups. A researcher uses the prior or extra information to create cut-points, or stratum boundaries for quickly and frugally separating the observations into groups. Generally, the researcher separates the observations into n strata using the cut-points. Following this, the researcher selects every k th observation from each stratum to measure using a costly measurement. An equal fraction of measurements comes from each stratum of observations. The final measured observations represent the range of values that appear across the n strata, while still forming a random sample.

2.3. Double sampling with ratio estimation

Double sampling with ratio estimation (for description see Thompson, 1992) is appropriate when the frugal measurement is a consistent, but less precise measurement of the desired costly measurement. In this case the relationship between the variable of interest, by the costly measurement, and the auxiliary variable, by the frugal measurement, is expressed as the slope of a linear regression of the data. In general, a number of observations are collected using the frugal measurement, and then a smaller number of the original sample is also observed using the more costly measurement. The samples observed twice, using both methods, are used to generate the linear regression, provided the model holds. The linear regression relationship can be exploited to obtain more precise estimates on the variable of interest than if only the small sample of costly measurements were observed.

3. DECIDING ON A SAMPLING METHODOLOGY BASED ON PRIOR KNOWLEDGE

Many statistical sampling methodologies have been proposed for the collection of physical samples, where the set of potential sampling sites cannot all be stratified in advance. The most basic method is simple random sampling. Beyond this, many methods are available in the literature. The four methods covered in this article are: simple random sampling (SRS), ranked set sampling (RSS), weighted double sampling with cut points (WDS) and double sampling with ratio estimation (RE). Deciding which of these designs is appropriate for one's study can be difficult for researchers. One criterion to use is the amount of prior knowledge available on the distribution of the variable being measured and its correlates (Table I). The best choice of a method depends upon how much prior information is available. The sampling methods are listed in ascending order of efficiency. The greater the prior knowledge available, the more efficient a method can be selected.

Simple random sampling (SRS) is performed by randomly choosing points from the population; no prior information regarding the distribution of the data is needed. Although simple to implement, the method can be problematic if measuring is costly or if locating samples is difficult. For example, locating randomly chosen points in a large field or forest can be time consuming or logistically infeasible. A number of authors (e.g. Stokes, 1977; Patil *et al.*, 1994; Johnson *et al.*, 1996) have suggested RSS as an alternative to SRS when measuring samples is expensive. RSS can result in greater precision than SRS while maintaining or reducing costs. Mode *et al.* (1999) extended this work by showing the minimal relative costs of measuring versus ranking required for RSS to have the same precision as SRS with equal total sampling costs. When ranking costs are non-trivial, it is possible for SRS to be preferable to RSS. This occurs when ranking does not increase the precision of the estimate enough to make it cost effective.

Table I. Prior knowledge required for successful use of sampling methods

Sampling method	Prior knowledge
Simple random sampling (SRS)	None
Ranked set sampling (RSS)	Frugal covariate for ranking
Weighted double sampling with cut points (WDS)	Frugal covariate, general distributional information for cut points
Double sampling using ratio estimation (RE)	Frugal measurement that is linearly related, and highly correlated with measured covariate at each sample point

Weighted double sampling (WDS) uses general information about the distribution of the data to address one of the potential problems in RSS. When ranking items in sets, it is possible to randomly include three very large items in a single set, such that the estimate for the first order statistic from that set is a very large value. Sampling would be more efficient if the data could be pre-stratified. WDS creates boundaries, or cut points, for the strata using prior knowledge of the distribution. If the cut points reasonably divide the distribution such that large percentages of the distribution fall into each stratum, then WDS will be preferable to RSS. No measured covariate is required, only the ability to place each sample into one of the strata defined by the cut points. If the cut points are not appropriate for the specific distribution (see Section 4.2, for example), then RSS can be preferable to WDS. In the extreme situation where all of the distribution falls into one of the WDS strata, WDS is equivalent to SRS except that it incurs stratification costs. In this situation SRS can even be preferable to WDS.

Double sampling using ratio estimation (RE) exploits a particular relationship between the frugal and costly information. The sampling model assumes that there is a regression-through-the-origin relationship between the auxiliary variable and the variable of interest. This relationship can be estimated through the measuring of several samples by both the frugal and costly measurement and used to 'correct' values evaluated only using the frugal measurement. RE is a powerful sampling design for environmental research when the error variance in the ratio estimation model is small compared to the variance in the costly measured values. This occurs when the auxiliary and main interest variables are highly correlated.

To summarize, if an informative, continuous, linearly related covariate exists, then one should use double sampling with ratio estimation. When such a covariate does not exist, but the information can be categorized, it is best to use weighted double sampling to produce the estimates. If there is little information on the covariate, then ranked set sampling is the best choice, assuming the relative cost of ranking to measuring is fairly small. If ranking is expensive relative to measuring, then simple random sampling is appropriate.

4. DEGREES OF PRIOR KNOWLEDGE: EXAMPLE

We used simulations from 21 empirical data distributions to examine the applicability of several sampling methods. (These data were also used in a paper examining cost efficiency of RSS to SRS: Mode *et al.*, 1999). The data sets consisted of visual ('measuring by eye') and physical (using measuring tapes) habitat unit length and width measurements from coastal Oregon streams. Figure 1 shows a plot of the measured stream habitat areas and visually estimated areas. Data were collected by the USDA Forest Service during stream monitoring inventories. The streams were mostly in forested areas which drain into the Pacific Ocean from the Umpqua River Basin north to the Columbia River Basin Boundary. Each stream had between 36 and 108 habitat units (median 50). The data consisted of habitat areas calculated from the visually and physically measured length and width of each habitat. Resampling with replacement was used to compare estimated means based on several sampling methods. Each stream was treated as an empirical distribution, and 4000 random, independent samples were drawn for each sampling method. The 21 streams, each with its own level of skewness (see Figure 2), were thus used to demonstrate the advantages and disadvantages of the different methods under a variety of distributions.

4.1. Methods

Four methods were examined: simple random sampling, ranked set sampling with a set size (n) of 3, weighted double sampling with three predetermined strata, and double sampling using a

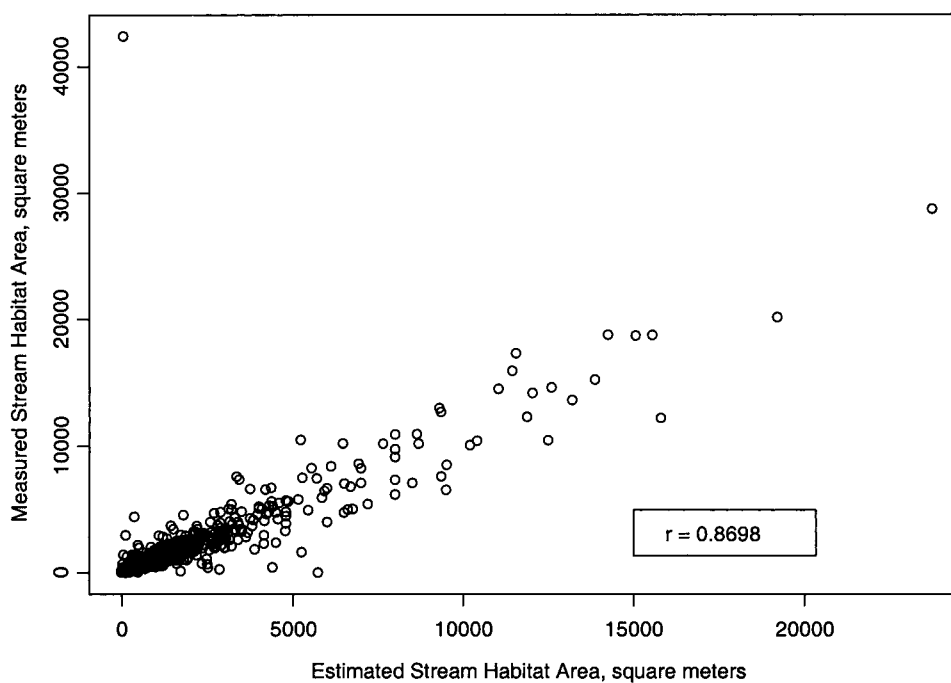


Figure 1. Plot of physically measured and visually estimated stream habitat areas

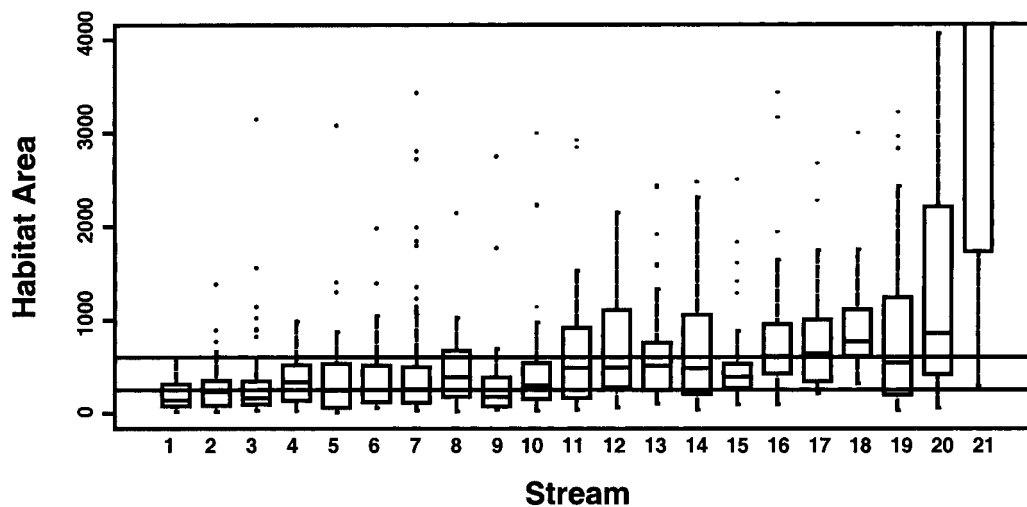


Figure 2. Distributions of actual habitat area for each of the 21 streams. Streams are ordered by their mean habitat area. Horizontal lines indicate the predetermined cut-off points for the weighted double sampling. The vertical axis has been truncated to aid in visualizing the cut points

ratio estimator. All methods used an average of 12 physically measured habitat areas to calculate the mean.

Simple random sampling (SRS) consisted of randomly selecting 12 habitat areas, and calculating the mean based on the physically measured habitat areas.

Ranked set sampling (RSS) consisted of randomly selecting a set of three habitats, ranking them according to the visual measurements, selecting the smallest area and recording the physical measurement. Then another set of three habitats were randomly selected, ranked according to the visual measurement, and the middle sized habitat area selected and the physical measurement recorded. Another set of three habitats were randomly selected, ranked, and then the largest habitat based on the visual measurement selected and the physical measurement recorded. This process was repeated four times, resulting in 12 physical habitat areas ($nm = 12$) to calculate the mean from the 36 areas observed ($n^2m = 36$).

Weighted double sampling (WDS) consisted of first preselecting strata cut-off points to be used with the visual measurements. Cut-offs were determined using data from a large data set of approximately 3000 stream habitats in Oregon. The cut-offs divided the data into small ($< 250 \text{ m}^2$), medium ($250\text{--}600 \text{ m}^2$) and large ($> 600 \text{ m}^2$) habitat areas (Figure 2, horizontal lines). Thirty-six habitats were randomly selected from the stream data and placed into one of the three strata based upon their visual measurement. Every third habitat was selected from each stratum beginning with the second one (i.e. 2nd habitat, 5th habitat, 8th habitat, ...) and the physically measured area from each was recorded. Since 36 habitats had been randomly selected, 11–13 physically measured areas were included in the final mean calculation. Each strata mean was weighted by the fraction of habitats observed in each stratum, the latter being an unbiased estimator of the fraction in each strata in the population. For example, if there were 14 large habitats, 16 medium habitats and 6 small habitats in a random sample, then the relative weights would be $14/36$, $16/36$ and $6/36$ for each stratum.

Double sampling using a ratio estimator (RE) consisted of randomly selecting 36 habitats and observing the visual measurements of habitat unit areas. From those habitats, 12 were randomly selected and the physically measured areas were also observed. The relationship between the physical and visual measurements was assumed to be a straight line through the origin. The ratio estimator was used to adjust the visual measurements (the slope of the line through the data) and to calculate the mean (see Cochran, 1977, Chapter 6). This was performed on each sample.

4.2. Results

It should be noted that the cut points from the larger sample of streams in Oregon divided most of the streams into three sections. For a few streams (see Figure 2, streams 20 and 21) the cut points were inappropriate and failed to adequately stratify the sample. Data from these streams were only placed into two strata, and thus some of the extra effort in identifying strata on a large number of samples was wasted. The improperly chosen cut points did not bias the sample, since all habitat areas were randomly chosen, but the extra effort employed did not result in greater precision of the estimate.

One way of comparing multiple sampling methods is to use relative precision (RP) as defined in survey sampling. For example, when comparing two methods, relative precision is defined as

$$RP = \frac{\text{var}(\bar{X}_i)}{\text{var}(\bar{X}_b)}.$$

The variances are of the sample mean calculated for a sample taken using each method. In general, RP is a ratio of the variance using the particular sampling method (*i*) relative to the 'base' or comparison method (*b*). This definition of relative precision is consistent with the concept of design effect used in survey sampling (Kish, 1965), but is the inverse of RP used in many ranked set sampling papers. For example, RPs of 2 and 0.75 indicate sampling methods with twice the variance and 75 per cent of the variance from a sample of the same size using the base method, respectively. The amount gained will be a function of the method and its appropriateness.

Relative precision values for each double sampling method versus SRS as a base were calculated for each stream (Figure 3). The values represented the variance in the means from 4000 resamplings. Overall the results followed what was expected. All three methods were more precise than SRS, with RE resulting in the lowest overall variance relative to SRS. WDS resulted in lower relative precision than RSS overall, but where the cut points were inappropriate (see Figure 2), RSS was more precise. Relative precision values found with WDS (range 0.47–0.89, mean 0.66) tended to be only slightly lower than those with RSS (range 0.56–0.95, mean 0.71) and may not have been worth the risk of using inappropriate cut points. This result demonstrates the importance of *good* prior information in establishing the strata cut points. When the cut points were appropriate for a stream, i.e. when the cut points separated the data into three groups, WDS was consistently better than RSS, reducing the variance by as much as 20 per cent. However, when the prior information was inappropriate, RSS was more precise than WDS. For example, in stream 21, almost all the habitat areas fell above the upper cut point, so WDS was almost equivalent to SRS. RE was generally successful, but failed in streams where the relationship between the visual and physical measurement was decidedly non linear. The original data were collected for use with RE, although actual calculations required more complex procedures (see Conquest *et al.*, 1991 for discussion of logistical and quality control issues).

5. CONCLUSIONS

The difficulties of environmental sampling can include expensive laboratory analyses and logistically cumbersome sampling frames. The desire of researchers to include their knowledge into the choice of samples is strong and reasonable, but doing so can bias results if done improperly. Double sampling methods including ranked set sampling, weighted double sampling, and double sampling with ratio estimation, provide efficient random sampling methods that include extra or prior information. The challenge to researchers is deciding which method is appropriate for the data and research situation.

Ranked set sampling, double sampling with cut points, and double sampling with ratio estimation can all provide increased precision and/or lower sampling costs over simple random sampling. The decision on the appropriate methodology to use for environmental sampling should be based on the prior information available on correlated auxiliary variables. If frugal measurements relative to actual analytic costs can be taken on a linearly related highly correlated auxiliary variable, then double sampling with ratio estimation is the preferred procedure. If such data are not available, but frugal measurements can be taken that allow for classifying the target variable into the general parts of its distribution, then weighted double sampling with cut points is preferred. If the general distribution is unknown, but frugal measurements on a correlated auxiliary variable can be taken, then ranked set sampling is appropriate.

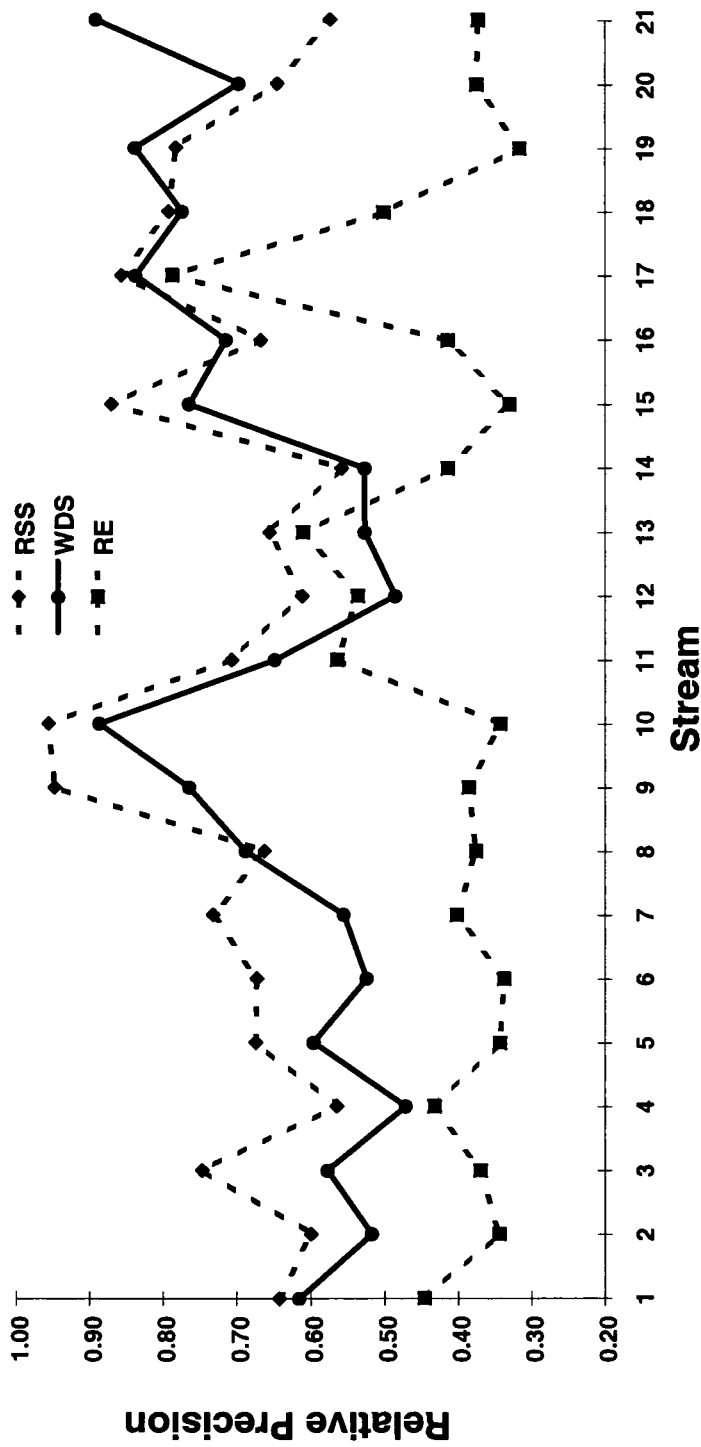


Figure 3. Relative precision values representing variance across 4000 resamplings for ranked set sampling (RSS), weighted double sampling (WDS), and double sampling using ratio estimation (RE), versus simple random sampling (SRS). Resampling was performed on each of the 21 Oregon stream data sets of habitat areas. Streams are ordered by their mean habitat area. Lower RP values indicate improvement versus using SRS

ACKNOWLEDGEMENTS

The authors would like to thank the National Research Center for Statistics and the Environment at the University of Washington for supporting this work, and Shaun McKinney of USDA Forest Service, Pacific Northwest Region for providing the stream data. Although the research described in this article has been funded in part by the U.S. Environmental Protection Agency through agreement CR825173-01-0 to the University of Washington, it has not been subjected to the Agency's required peer and policy review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred. L. L. Conquest received additional support from the U.S. Environmental Protection Agency's National Health and Environmental Effects Research Laboratory, Western Ecology Division, through cooperative agreement CR824682. The authors would also like to thank Graham Kalton, Westat, for his suggestions concerning double sampling, and Rebecca Buchanan (University of Washington) for additional technical and data analysis assistance.

REFERENCES

- Cochran WG. 1977. *Sampling Techniques*. John Wiley & Sons: New York.
- Conquest LL, Cardoso TP, Seidel KD, Ralph SC. 1991. Using visual estimation in a watershed monitoring study. Appendix C. *Status and Trends of Instream Habitat in Forested Lands of Washington: The Timber-Fish-Wildlife Ambient Monitoring Program*. Biennial Progress Report, Center for Streamside Studies. Seattle: Washington.
- Johnson GD, Nussbaum BD, Patil GP, Ross PN. 1996. Designing cost-effective environmental sampling using concomitant information. *Chance* **9**(1): 4–11.
- Kish L. 1965. *Survey Sampling*. John Wiley & Sons: New York.
- McIntyre GA. 1952. A method for unbiased selective sampling, using ranked sets. *Australian Journal of Agricultural Research* **3**: 385–390.
- Mode NA, Conquest LL, Marker DA. 1999. Ranked set sampling for ecological research: accounting for the total costs of sampling. *Environmetrics* **10**(2): 179–194.
- Patil GP, Sinha AK, Taillie C. 1994. Ranked set sampling. In *Handbook of Statistics* (pp. 167–200), Patil GP, Rao CR (eds). North-Holland: New York.
- Stokes SL. 1977. Ranked set sampling with concomitant variables. *Communications in Statistics—Theory and Methods* **A6**: 1207–1211.
- Takahasi K, Wakimoto K. 1968. On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics* **20**: 1–31.
- Thompson SK. 1992. *Sampling*. John Wiley & Sons: New York.